Large Language Models are not Fair Evaluators

Year: 2023 | Citations: 759 | Authors: Peiyi Wang, Lei Li, Liang Chen

---

Abstract

In this paper, we uncover a systematic bias in the evaluation paradigm of adopting large language models~(LLMs), e.g., GPT-4, as a referee to score and compare the quality of responses generated by candidate models. We find that the quality ranking of candidate responses can be easily hacked by simply altering their order of appearance in the context. This manipulation allows us to skew the evaluation result, making one model appear considerably superior to the other, e.g., Vicuna-13B could beat ChatGPT on 66 over 80 tested queries with ChatGPT as an evaluator. To address this issue, we propose a calibration framework with three simple yet effective strategies: 1) Multiple Evidence Calibration, which requires the evaluator model to generate multiple evaluation evidence before assigning ratings; 2) Balanced Position Calibration, which aggregates results across various orders to determine the final score; 3) Human-in-the-Loop Calibration, which introduces a balanced position diversity entropy to measure the difficulty of each example and seeks human assistance when needed. We also manually annotate the"win/tie/lose"outcomes of responses from ChatGPT and Vicuna-13B in the Vicuna Benchmark's question prompt, and extensive experiments demonstrate that our approach successfully mitigates evaluation bias, resulting in closer alignment with human judgments. We release our code and human annotation at \url{https://github.com/i-Eval/FairEval} to facilitate future research.